

# Targeting treatments for depression: what can our patients tell us?

Received 7 March 2016; Accepted 8 March 2016; First published online 5 April 2016

**Key words:** Depression, treatment selection, clinical indicators.

Commentary on: Kessler *et al.* (2016). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*. (doi: 10.1017/S2045796016000020)

There is no dispute: there is great heterogeneity in the treatment effects associated with major depressive disorder (MDD). Since depression is a syndrome, we would expect a large range of causes, potential treatments and outcomes with any one treatment as well as between treatments. As Kessler's report notes, the field has yet to effectively address and reduce this heterogeneity despite multi-decade attempts including descriptively defined depressive subtypes (Rush *et al.* 2008) and – more recently – neuropsychological (Etkin *et al.* 2015), imaging (Korgaonkar *et al.* 2014) and genetic (Schatzberg *et al.* 2015) approaches.

Kessler and colleagues review a variety of recent efforts and methodological innovations to better address the challenge of the heterogeneity of treatment effects (HTE) in MDD. They have compiled a list of critical baseline patient-reported parameters that have at least some evidence of relating to outcome in the acute treatment of depression. These 30+ measures are easily obtained from patients at very low cost, although the optimal combination for predicting treatment outcomes or selecting between treatments is yet undefined since the full compendium of measures has not been tested in large informative populations of depressed patients.

The authors properly emphasise the extraordinary importance of such an effort. Indeed, the relevant combination of these inexpensively acquired measures can and should provide a platform upon which additional laboratory-based measures should be placed to further reduce HTE and better target treatments. In fact, a lack

of knowledge regarding what could be offered by these patient-reported indicators could limit the clinical value of specific laboratory tests. This idea is not surprising because clinical parameters typically influence whether treatments will work.

Finally, the authors provide an important demonstration of an analytic approach using epidemiological data. Their proposal clearly deserves testing in large informative samples on which all the suspected baseline predictors are collected. Large samples, combined with machine learning and replication of the initial results in independent samples, seems very likely to yield clinically actionable information and clinically useful tools that should inform clinical decision-making.

There are, however, some challenges and questions to be considered in taking this proposal further. One challenge is to define the preferred sample. While one may hope that machine learning will sort out many issues if samples are big enough, unnecessarily large samples may add more cost and complexity than benefit. At first glance, would not we want to include all patients who are deemed to be clinically depressed and are receiving medication? Extremely inclusive samples may not be more informative, especially if: (a) our treatments are not uniquely effective in distinct subsamples, and (b) the accuracy of the chart diagnosis is questionable. Both seemingly apply.

We know, for example, that some depressed patients suffer from large numbers of comorbid general medical conditions that in turn affect the efficacy of treatments (Rush *et al.* 2006, 2009; Rush, 2007). If a medication is ineffective in some substantial subset of depressed patients, the case mix under study may well affect the algorithms being developed, which could reduce the chances of replication in another independent population that has a different case mix. While machine learning may overcome some of these issues to some degree, a more cost-effective approach may entail fewer subjects, which increases chances for replication in independent samples. Relying on

---

Address for correspondence: A. J. Rush, Duke-National University of Singapore, 6, The Academia, 20 College Road, 169856, Singapore. (Email: [john.rush@duke-nus.edu.sg](mailto:john.rush@duke-nus.edu.sg))

randomised controlled trials that are designed for maximal internal validity limits subject availability and leads to unrepresentative samples that do not represent patients in practice, making results potentially challenging to replicate (Wisniewski *et al.* 2009).

Another sample selection issue is whether to include treatment-resistant with multiple prior failed treatment attempts. Such a sample is likely to be less heterogeneous compared with all patients who entered the first treatment step, and placebo response rates should be lower.

If the placebo responders are common in the sample (logically more likely with the first than the second treatment step), they may mask the detection of indicators that reduce heterogeneity among those who do respond. While very large samples may help with this challenge, there are costs.

Furthermore, if a combination of inexpensively acquired patient-reported variables actually reduces HTE, then that particular combination should become a platform upon which more expensive laboratory tests can be added. Developing the combination of patient-reported variables that best informs the first step would certainly be useful. However, the addition of laboratory tests will likely be more cost-effective at the second or subsequent steps. Thus, a second platform using perhaps a different or even the same combinations of patient-reported variables would be of value in informing this second step in anticipation of the addition of laboratory tests. It would seem important to define the clinical decision points (which steps and concerning which choices) to be addressed by the results of machine learning.

This could be accomplished in a registry in which patients begin with a single generic medication, after which the second step could offer some common choices. Such a registry would be readily accepted by patients, even if one limited the number of second step options (to focus on specific clinically important questions or common decisions). An equipoise-stratified randomised design could be a consideration (Lavori *et al.* 2001).

Another truly major challenge entails how to minimise noise generated by wide variation in treatment delivery. Antidepressant medications are underdosed, patient adherence is highly variable, clinician adherence to practice clinical guidelines is quite variable, and up to one-third of patients do not complete acute medication trials. Measurement-based care does produce more robust but tailored dosing and better outcomes than routine care (Rush, 2015), but measurement-based care is not yet widely used. Large samples of improperly treated patients could produce machine learning results that do not apply to higher-quality treatment practices. Inappropriate

heterogeneity in care delivery simply adds to HTE, and lowers the chances of replication in independent samples with higher-quality care.

Another challenge is the absence of any measured outcome (e.g., depressive symptoms or function) in most electronic health records (EHRs). In this context, we would not know how well each patient has fared. If measurement-based care were implemented, it would have the dual advantage of improving the quality of care and providing a clinically relevant outcome that could be entered into the EHR. Even if symptom ratings were used by different providers, such as the Montgomery Asberg Depression Rating Scale, the Patient Health Questionnaire or the Quick Inventory of Depressive Symptoms ([www.ids-qids.org](http://www.ids-qids.org)), item response theory analyses can provide reasonably good crosswalks between these various measures ([www.ids-qids.org](http://www.ids-qids.org)).

Two final thoughts: First, let us not overlook the potential value of combining baseline patient self-reported variables with early post-baseline changes in either symptoms or function to target treatments. A *post hoc* analysis of STAR-D (Rush *et al.* 2004) data provided a proof of concept of this approach (Kuk *et al.* 2010; Li *et al.* 2012). Machine learning in large samples with both baseline and early post-baseline data would clearly strengthen the clinical utility and precision of such an approach. These kinds of results could provide an evidence-based approach to address another clinical challenge; namely, to avoid prolonged treatment trials that are certain to produce a poor outcome. This challenge should be addressed as soon as possible using baseline and early post-baseline patient-reported parameters.

Second, these patient-reported baseline parameters may also sub-serve other clinical functions beyond reducing HTE to better target treatments. Clinically available algorithms developed by machine learning could help to predict longer-term prognoses for depressed patients following response or remission. For example, could we identify the 10–18% of remitted depressed patients who are likely to relapse within a year (Judd *et al.* 2015) using a combination of either baseline or end-of-acute-treatment self-reports? If so, follow-up visit frequency could be tailored to individuals based on their likelihood of relapse.

In summary, a combination of self-reported baseline parameters would appear to be feasible, inexpensive and very likely to reduce HTE and thus enhance treatment targeting. This tool is also likely to provide an essential platform upon which more time-intensive and expensive tests to further reduce HTE could be evaluated and developed. A multi-site registry would seem to be essential to ensure a reasonably representative patient sample, perhaps focused on a select number of treatment options and specific treatment steps,

all delivered by measurement-based care. This effort would be one way to increase feasibility and contain cost. Potentially different registries that are focused on clinically important but distinct subgroups (e.g., bipolar and unipolar patients; depressed youth; adults; elderly) could be developed and similarly analysed. Given the current remarkable dearth of clinically informative tools to reduce HTE, the table is set for substantial advances.

**A. John Rush**

*Duke-National University of Singapore, Singapore*

## References

- Etkin A, Patenaude B, Song YJC, Usherwood T, Rekshan W, Schatzberg AF, Rush AJ, Williams LM** (2015). A cognitive-emotional biomarker for predicting remission with antidepressant medications: a report from the iSPOT-D trial. *Neuropsychopharmacology* **40**, 1332–1342.
- Judd LL, Schettler PJ, Rush AJ, Coryell WH, Fiedorowicz JG, Solomon DA** (2015). A new empirical definition of major depressive episode recovery and its positive impact on future course of illness. *Journal of Clinical Psychiatry* (epub ahead of print). doi: 10.4088/JCP.15m09918.
- Korgaonkar MS, Rekshan W, Gordon E, Rush AJ, Williams LM, Blasey C, Grieve SM** (2014). Magnetic resonance imaging measures of brain structure to predict antidepressant treatment outcome in major depressive disorder. *EBioMedicine* **2**, 37–45. doi: 10.1016/j.ebiom.2014.12.002.
- Kuk A, Li J, Rush AJ** (2010). Recursive subsetting: a method to enhance the accuracy of early prediction of treatment outcome and to inform personalized care. *Journal of Clinical Psychiatry* **71**, 1502–1508.
- Lavori PW, Rush AJ, Wisniewski S, Alpert J, Fava M, Kupfer DJ, Nierenberg A, Quitkin FM, Sackeim HA, Thase ME, Trivedi M** (2001). Strengthening clinical effectiveness trials: equipoise-stratified randomization. *Biological Psychiatry* **50**, 792–801.
- Li J, Kuk A, Rush AJ** (2012). A practical approach to the early identification of antidepressant medication non-responders. *Psychological Medicine* **42**, 309–316.
- Rush AJ** (2007). STAR\*D: what have we learned? [special article]. *American Journal of Psychiatry* **164**, 201–204.
- Rush AJ** (2015). Isn't it about time to employ measurement-based Care in practice? *American Journal of Psychiatry* **172**, 934–936.
- Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM, Kupfer DJ, Rosenbaum JF, Alpert J, Stewart J, McGrath PJ, Biggs MM, Shores-Wilson K, Lebowitz BD, Ritz L, Niederehe G, for the STAR\*D Investigators Group** (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR\*D): rationale and design. *Controlled Clinical Trials* **25**, 119–142.
- Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, Niederehe G, Thase ME, Lavori PW, Lebowitz BD, McGrath PJ, Rosenbaum JF, Sackeim HA, Kupfer DJ, Luther J, Fava M** (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR\*D report. *American Journal of Psychiatry* **163**, 1905–1917.
- Rush AJ, Wisniewski SR, Warden D, Luther JF, Davis LL, Fava M, Nierenberg AA, Trivedi MH** (2008). Selecting among second-step antidepressant medication monotherapies. Predictive value of clinical, demographic, or first-step treatment features. *Archives of General Psychiatry* **65**, 870–881.
- Rush AJ, Warden D, Wisniewski SR, Fava M, Trivedi MH, Gaynes BN, Nierenberg AA** (2009). STAR\*D: revising conventional wisdom. *CNS Drugs* **23**, 627–647.
- Schatzberg AF, DeBattista C, Lazzeroni LC, Etkin A, Murphy GM Jr., Williams LM** (2015). ABCB1 Genetic effects on antidepressant outcomes: a report from the i-SPOT-D. *American Journal of Psychiatry* **172**, 751–759.
- Wisniewski SR, Rush AJ, Nierenberg AA, Gaynes BN, Warden D, Luther JF, McGrath PJ, Lavori PW, Thase ME, Fava M, Trivedi MH** (2009). Can Phase III trial results of antidepressant medications be generalized to clinical practice? A STAR\*D report. *American Journal of Psychiatry* **166**, 599–607.